MORAL COGNITION AND ITS NEURAL CORRELATES:
POSSIBILITIES FOR ENHANCEMENT OF MORAL COGNITION AND BEHAVIOR

Bachelor Degree Project in Cognitive Neuroscience
Basic level 22.5 ECTS
Spring term 2018

Elin Vidlund
Supervisor: Stefan Berglund
Examiner: Sakari Kallio
Abstract

This essay aims to provide an overview of some key theories and frameworks regarding moral cognition and its neural correlates, in order to examine the possibilities of enhancement of moral cognition. Moral cognition arises from the functional integration of several distinct brain regions and networks. These neural systems correspond to different socioaffective abilities, such as empathy and compassion, as well as sociocognitive abilities, such as theory of mind. Due to this neural distinction, these moral abilities, behaviors, and emotions can be targeted and trained separately. Recent research suggests that training sociocognitive and socioaffective abilities increases cortical thickness in corresponding brain regions and networks, hence providing support for adult neural plasticity in relation to moral cognition. Increased cortical thickness also corresponds to enhanced performance in socioaffective and sociocognitive abilities. Training compassion and empathy induce enhanced abilities to pick up emotional cues, as well as strengthen the motivation to alleviate others’ distress. Practicing theory of mind allows for a better understanding of the perspective of others, which has been indicated to reduce biases between individuals or societal groups. Thus, enhanced moral cognition can contribute to an increase in consideration for those affected by our choices and behavior, which may yield more compassionate, just, and safe societies.

Keywords: Moral cognition, moral enhancement, neural correlates of moral cognition, empathy, compassion, theory of mind.
# Table of Contents

Introduction ................................................................................................................................. 3

What is Morality? .......................................................................................................................... 5

  Key Concepts of Morality ......................................................................................................... 6

  The Evolution of Morality ......................................................................................................... 9

  Moral Dilemmas and Moral Judgements ................................................................................. 11

The Neural Correlates of Moral Cognition ................................................................................. 13

  The Interaction of Cognitive Control and Affective Processing .............................................. 14

  Moral dilemmas: Utilitarian and deontological judgements ..................................................... 15

  Altruism, cooperation and pro-social behavior ........................................................................ 17

  Empathy and compassion ......................................................................................................... 18

  Perspective taking and theory of mind .................................................................................... 21

  The dual-process theory ........................................................................................................... 22

Moral Enhancement ..................................................................................................................... 25

  Socioaffective Training ............................................................................................................ 26

  Sociocognitive Training ........................................................................................................... 29

  Sociocognitive and Socioaffective Training with Interactive Aspects ..................................... 30

Discussion .................................................................................................................................. 35

References ................................................................................................................................... 40
Introduction

Imagine standing next to a train track on which five people are trapped. A trolley is heading towards them and the only way to save their lives is by hitting a switch turning the trolley onto another track. However, on this other track there is also one person trapped and will be killed if the trolley is turned over. Ought you to save five people forfeiting one by hitting the switch? Now, imagine another similar scenario in which you are standing on a footbridge over the tracks next to a large stranger. The oncoming trolley threatens to kill the same five people and now the only way to save them is by pushing the large stranger onto the tracks, and thus stopping the trolley but killing the stranger. Ought you to push a person to their death in order to save five others (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001)?

These are examples of moral dilemmas which may allow for a greater understanding of moral cognition (i.e. how humans process morally salient information). In regard to these two moral dilemmas there is a general common consensus. In the first scenario most people would say yes, whereas in the second scenario most people would say no (Greene et al., 2001). These conflicting answers to different, yet very similar, moral dilemmas can aid in the understanding of how human moral cognition arises, as well as what different factors make up human moral cognition and what may affect it. In today's modern society we may stand before rather sophisticated and abstract moral problems on a daily basis. Hence, with greater knowledge of how the human brain responds to moral situations we are better equipped to make good moral decisions.

Recent research investigates how moral cognition arise from the coordinated interaction of various neural systems (Moll et al., 2005). These include systems that enable reasoning and cognitive control (Zahm, 2006), empathy and compassion (Decety & Cowell, 2015; Lamm, Decety, & Singer, 2011; Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman,
2005), and the ability to take on the perspective of others regarding their beliefs, thoughts, and emotions, which is usually referred to as having a theory of mind (ToM) (Shamay-Tsoory, Tomer, Berger, Goldsher, & Aharon-Peretz, 2005).

Moreover, recent research also investigates the possibilities of moral enhancement, which involve the improvement of sociocognitive abilities, such as ToM, and socioaffective abilities, such as empathy and compassion, to foster pro-social behavior. These abilities have been indicted to be enabled by separate neural systems and thus can be targeted, trained, and strengthened separately (Kanske, Böckler, Trautwein, & Singer, 2015). According to Valk et al. (2017), training sociocognitive and socioaffective abilities enhance activity and integration of the brain regions and networks commonly associated with these abilities, as well as improve pro-social performance. Hence, training induced enhancement of moral cognition has been indicated to be a possibility (Valk et al., 2017), which may render interpersonal benefits such as reduced bias between different social groups (Kang, Gray, & Dovidio, 2014; Oatley, 2016). Thus, enhanced moral cognition can contribute to an increase in consideration for those affected by our choices and behavior, as well as increased willingness to relieve others’ distress, which may yield a more compassionate, equal, just, and safe society.

The aim of this essay is to give an overview of some of the main theories and frameworks regarding moral cognition, as well as the neural correlates of moral cognition, in order to examine the possibilities of moral enhancement. In this pursuit, this essay will firstly provide an overview of some central philosophical and psychological theories, concepts, and frameworks regarding moral cognition. Secondly, there will be an overview of the current neural correlates of moral cognition, centered around these key theories, concepts, and frameworks. Subsequently, there will be an in-depth critical study of the most recent research and prominent theories regarding the possibilities enhancement of moral cognition and behavior. Toward the end of this essay there will be a discussion regarding the previous
chapters and its theoretical and real-life implications, including a conclusion where the main points of the essay will be summarized. This essay will only consider moral cognition and its neural correlates in healthy adults, and thus will not cover the development and possible enhancement of moral cognition in children. Furthermore, strategies for enhancement of moral cognition and behavior will not be presented as treatments of any mental illnesses or impairments in unhealthy or abnormal functioning adults. This essay will only cover impairments of moral cognition in the pursuit of explaining normal function.

What is Morality?

The evolution of cooperation and pro-social behavior need not have been driven by the prospect of selfish gains, but by the enhanced survival prospects for cooperative human groups (Gintis, Henrich, Bowles, Boyd, & Fehr, 2008). Conversely, cooperation may also easily be exploited by one individual for their own selfish gains given the possibility (Rand et al., 2014). Hence, according to Haidt and Kesebir (2010), the function of morality is to regulate selfish behavior and to foster cooperation and pro-social behavior (Haidt & Kesebir, 2010).

Morality has many different propositional models and theories which put varying emphasis on the different cognitive and affective abilities that constitute human moral cognition. In this section, two of the main models regarding moral cognition will be presented (the rationalist model and the social intuitionist model) in order to highlight this varying importance given to cognitive and affective abilities. In the following subsections, some of the main sociocognitive (theory of mind) and socioaffective aspects (empathy, compassion, altruism, sympathy/empathic concern) of moral cognition will be addressed, as well as the three main theories and frameworks on solving moral dilemmas (by deontological or utilitarian judgments, as well as the dual-process theory).
The rationalist model (Haidt, 2001) states that moral judgements are made possible by a process of reasoning and reflection. Moral emotions, such as compassion and empathy, play only a small role as inputs in this rational process. Instead, one weighs issues of harm, rights, justice and fairness before passing a judgement whether something is morally right or wrong. Notably, according to Haidt and Kesebir (2010), the cognitive aspects of moral cognition (such as ToM) can also be rather intuitive and automatic, and do not necessarily rely on laborious rational efforts.

The social intuitionist model (Haidt, 2001), which is based on the ethical intuitionist view that there are certain “moral truths” (Haidt, 2001, p. 814) that are intuitive and come without a process of ratiocination. The social aspect of the social intuitionist model proposes that moral reasoning is usually a process which occur after the event with which one tries to justify one’s intuitions and emotions or influence on others’ intuitions and emotions, thus passing moral judgment is an interpersonal process (Haidt, 2001).

**Key Concepts of Morality**

The common denominator between these previously mentioned theories and models regarding moral cognition is that it consists of affective and cognitive aspects, which are all considered to be intertwined to various extents. This section will present the most commonly associated sociocognitive and socioaffective aspects of moral cognition. As such, the sociocognitive aspect (ToM), as well as the socioaffective aspects (sympathy/empathic concern, altruism, compassion, and lastly empathy), will be briefly addressed according to proposed definitions. However, compassion and empathy could be argued to be two of the most prominent aspects of moral cognition. Hence, compassion and empathy will be discussed in further detail, including addressing the three different constituting components of these two aspects.
ToM refers to our ability of making abstract inferences regarding other peoples’ thoughts and emotions, and thereby understand and predict their behavior and intentions. This allows us to navigate very complex social relationships, and also aids in empathic responses to others (Shamay-Tsoory, et al., 2005).

Sympathy refers to the awareness or recognition of someone else’s distress, and the experience of feeling concern for their well-being (Escalas & Stern, 2003). Similarly, according to Gleichgerrcht and Young (2013), empathic concern refers to the kind feelings and experience of concern which may arise when witnessing another individuals’ distress.

Altruism, or altruistic behavior, is an action performed for the benefit of another, often at the expense of one self. (Neff & Seppala, 2016).

Compassion usually refers to the kind and caring emotions which arise as a response to the perceived suffering of others, which in turn evokes a desire to help and relieve the plight (Goetz, Keltner, & Simon-Tomas, 2010; Neff & Seppala, 2016). Controversially, according to Feldman (2005), compassion also entails a cognitive aspect of nonjudgmental understanding and recognition of shared human fallibility (Feldman, 2005). In addition to the evident pro-social aspects of compassion, it has also been positively linked to emotional self-efficacy, as well as the stress-reducing parasympathetic autonomic nervous system (Goetz, Keltner, & Simon-Tomas, 2010). As such, compassion may render benefits at an individual level, as well as at interpersonal and societal levels. Furthermore, according to Moll et al. (2005), compassion arises from the integration of three components, the first being perceptual cues that does not rely on context (such as a child’s distressed facial expression). The second being abstract knowledge regarding social concepts (such as the vulnerability of an orphan). Lastly, compassion also arises from evoked central motive states (such as distress, anxiety, and social attachment) with a body of information regarding the context of the situation (such as the underlying tragic cause of the child becoming an orphan, and that the possibility of
MORAL COGNITION AND ITS NEURAL CORRELATES: POSSIBILITIES FOR ENHANCEMENT OF MORAL COGNITION AND BEHAVIOR

adoption might be slim). Moll et al (2015) goes on to explain that central motive states instigate helping behavior by eliciting affective states (such as distress, anxiety, social attachment, hunger, and happiness), which are not directed to any specific danger or critical situation. These affective states are distinguished from basic emotions, such as fear, which arise from the integration of contextual perception (i.e. directed to a feared object or situation) and central motive states (Moll et al., 2005).

Empathy usually refers to the ability to acknowledge- and subjectively experience similar feelings as those expressed by others. The term usually also includes the aspect of not confusing the source of the experienced feelings, and hence be able to separate the expressions generated by oneself or others (Decety & Jackson, 2006). This ability to recognize and model feelings and behavior (such as detecting distress in group members) is not unique to humans (Decety, 2011a, 2011b). Non-human empathy has many different layers with varying complexity. At its most emergent stage non-human empathy involves motor mimicry and mirrored neural representations, and progresses into more advanced abilities such as ToM in large-brained species (De Waal, 2012).

The definition of empathy has been expressed in multiple ways based on different criteria (Batson, 2009). Thus, making it difficult to differentiate between which psychological process or mental state that is referred to, which can vary in their biological functions and neural representations. This in turn also affect the relationship between empathy and moral behavior (Decety & Cowell, 2015). Although, there is a broad consensus that empathy is a core ingredient of moral cognition (Decety & Jackson, 2006), the experience of moral thinking can occur without empathy (Haidt & Joseph, 2007).

According to Decety and Cowell (2015), empathy consist of three components which constitute the three pillars of empathy; the affective component, the motivational component, and the cognitive component, which each has a different relationship with moral cognition.
The affective component of empathy consists of emotional contagion or affective resonance, which refers to the ability to share or become affectively aroused by others’ emotions. The motivational component of empathy shares similarities with compassion and reflects the willingness and compulsion to care for another individual’s welfare. Lastly, the cognitive component of empathy is similar to ToM and refers to the capacity to put oneself into the mind of another individual and imagine what that person is thinking or feeling from their point of view (Decety & Cowell, 2015).

Moreover, according to Moll et al. (2005), moral emotions are correlated with different variations of social and cultural norms, values, and contexts. Conversely, Moll et al. (2007) go on to explain that moral emotions are represented across cultures (Moll et al., 2007), and has also been observed in other primates (De Waal, 2012). In line with this, Nichols (2002) explains that the ability to make moral judgements about normative violations is made possible by the interaction of an evoked affective response as well as knowledge about current societal norms.

**The Evolution of Morality**

According to Marazziti, Baroni, Landi, Ceresoli, and Dell’Osso (2013), the moral sense is innate in humans, and has according to Gintis et al. (2008) developed from the benefits of altruism and pro-social behavior to the survival of individuals as well as group-constellations. Thus, rendering a competitive advantage over groups constituted of selfish individuals (Gintis et al., 2008). The evolutionary benefits of these altruistic and cooperative actions have enabled humans to excel above any other species in regards of reciprocity and a sense of fairness (Nowak & Sigmund, 2005). The evolution of the human sense of fairness can be further understood by observing other cooperative species, such as dogs, crows, and macaques. These species have displayed a sense of fairness in its emergent stages, such as reacting negatively when receiving a lesser reward for the same work as a conspecific.
Furthermore, genetically closer non-human relatives (i.e. chimpanzees) have demonstrated a more advanced sense of fairness by displaying aversion when a conspecific receives a lesser reward compared to their own better reward when performing the same task. Hence, showcasing a reaction for inequity regardless of the opportunity of selfish gains (Brosnan & De Waal, 2014). The development of a sense of fairness and aversion to inequity are according to Brosnan and De Waal (2014) a prominent contributor to the exceptional relative brain enlargement of the human species. Similarly, Holloway (2008) also argues that the human brain’s relative increase in volume and its current internal reorganization is primarily due to the social aspects of human development and the complexity that entail. This is in line with the social domain theory which postulates that human morality stems from the problems social interactions often evoke, such as conflicts, emotional hurt, physical violence, social injustices (Turiel, 2014).

Socialization has contributed to humans, like other social species, developing other dispositional characteristics, such as empathy, suitable for group living (Decety & Cowell, 2015). However, this social evolution has also given us some disadvantages, such as tribalism, which is the active favoring of members of one’s own group or tribe (i.e. ingroup member) versus members of another perceived group or tribe (i.e. outgroup member) (Greene, 2017). Similarly, empathy, which is usually a highly regarded moral emotion, can also be a source of immoral behavior. This since empathy has been shown to make people more prone to favor perceived ingroup members in resource-allocation situations (Batson, Klein, Highberger, & Shaw, 1995). However, this partiality may not be due to any genetic relatedness alone, but rather due to previous social experience and familiarity. This since it has been found that rats exhibit pro-social behavior to others only if they have previous social experience of that particular strain, regardless of genetic relatedness (Ben-Ami Bartal, Rodgers, Bernardez Sarria, Decety, & Mason, 2014).
Greene (2015), goes on to explain that humans possess psychological features that are distinctly moral, such as empathy, compassion, ToM, guilt, and aversion to violence. However, humans also have psychological features that are not so straightforwardly moral, such as gossip, ingroup favoritism, embarrassment, and vengefulness, which could all be argued to potentially be harmful or uncomfortable to experience. These combined moral psychological features allow inherently selfish individuals to gain benefits from cooperation and pro-social behavior. This in turn give us the ability to achieve goals which would not be possible with collective selfishness, and hence enable us to collectively live a better life just because we help and care for each other (Greene, 2015).

Moral Dilemmas and Moral Judgements

A moral dilemma is a hypothetical real-life situation which contain no obvious right or wrong answer or decision (Crockett, 2013). Moral dilemmas are useful when studying moral neuroscience because they pit competing processes against one another and are therefore high-contrast stimuli. Thus, moral dilemmas can reveal the cognitive structures that are engaged in moral decision-making (Cushman & Greene, 2012).

The concept of moral dilemmas can be subdivided into personal- and impersonal dilemmas (Greene et al., 2001). Personal dilemmas are considered to be situations in which the action in question requires direct or physical contact or aggression towards another human being (i.e. the footbridge dilemma). Impersonal dilemmas are considered to involve committing a harmful action towards an individual indirectly as a concomitant of another action (i.e. the trolley problem) (Greene et al., 2001).

Moral dilemmas typically evoke different strategies of solving them, usually based on three main schools of thought in normative ethics; consequentialism, deontology, and virtue ethics (Hursthouse & Pettigrove, 2016), or the dual-process theory (Greene et al., 2008).
Consequentialism views the outcome and consequence of an action as the sole measure for the moral acceptability of the action. As a part of the consequentialist view there is also the more specific view of utilitarianism. Utilitarianism states that consequences of an action should be assessed in terms of the maximization of well-being and reduction of suffering for as many people as possible (Greene, 2017). According to Gleischgerrcht and Young (2013), increased tendencies for utilitarian judgements have been associated with reduced dispositional empathic concern.

Deontology emphasizes rules or duties as the base upon which the moral acceptability of an action can be determined. Thus, unlike consequentialism, some actions or choices cannot be morally justified by their outcomes, only by the rules or duties that is set as a moral norm. Hence, an action or choice which render morally good consequences can be deemed as immoral, if the action or choice itself is morally forbidden (Alexander & Moore, 2016).

Virtue ethics emphasizes the importance of virtues or moral character in moral judgements. This is not to say that virtues are exclusive to the theory of virtue ethics. In consequentialism virtues are defined as traits that render favorable consequences, and in deontology virtues are defined as traits possessed by those who are loyal to the fulfillment of their duties. What distinguishes virtue ethics from other theories is the emphasis put in virtue within the theory and the resistance to define virtues as merely a part of another concept that is considered to be more fundamental (Hursthouse & Pettigrove, 2016).

The dual-process theory is proposed to be the integration of emotional and cognitive processes in response to a moral dilemma (Greene et al., 2008). According to Greene (2015), there is no specific moral faculty in the human brain which is used to make moral judgments. Instead, moral cognition arises from the functional integration of multiple neural mechanisms and cognitive systems, which each on their own is not specifically dedicated to moral judgment. Moreover, Greene, Morelli, Lowenberg, Nystrom, and Cohen (2008) go on to
explain that when we as humans are faced with a moral dilemma there are to be two thought-processes that guide our response. One is a fast and automatic response-process, which is an emotional reaction usually called a “gut reaction”. The second is a slower and more controlled process that is guided by reason and self-control (Greene et al., 2008). It is the integration of these two thought-processes that is the basis of the dual-process theory (Greene et al., 2008). In contrast, Prinz (2015) argues that this alleged regulation of emotional responses by recruiting reason is not due to reason suppressing or overriding of emotional responses, but rather to the recruitment of more neurological pathways connected to emotions. Thus, according to Prinz (2015), moral judgements are mainly based on emotions. However, the dual-process theory has been supported by Haidt (2001) who claim that moral judgements rely on moral intuitions which are automatic and effortless in their nature. Any justifications for these judgements rely on a subsequent and consciously laborious process (Haidt, 2001), hence suggesting that there is an initial fast- and automatic affective response to a moral dilemma, followed by a slower rational process. According to Crockett (2013), this slower rational form of moral decision-making is usually elicited when one is faced with a moral dilemma.

Lastly, there is indications that many moral judgements are highly uniform despite demographic differences such as ethnicity, gender, age and religious convictions (Young & Saxe, 2011). Hence, suggesting an objectivist view of moral judgement (Decety & Cowell, 2015).

The Neural Correlates of Moral Cognition

The neural organization of prosocial behavior is extensively more complex than that which underpins selfish behavior. This complexity is a result of the subtle and often swiftly changing mental states that underpin moral cognition (De Oliveira-Souza, Zahn, & Moll,
As previously mentioned, current neuroscientific research supports the notion that moral cognition is not a product of a dedicated moral faculty in the human brain (Greene, 2015). Hence, according Moll et al. (2005), the neural circuits involved in moral cognition are better to be regarded as a synchronized activity of several subordinate neuropsychological components. This section will address the most commonly mentioned brain regions and interactive networks associated with moral cognition.

Firstly, there will be an overview of the interaction between the cognitive and affective aspects of moral cognition, which will also be present through all following subsections regarding the neural correlates of utilitarian and deontological judgements, altruism, cooperation and pro-social behavior, empathy and compassion, theory of mind, and lastly the dual-process theory.

**The Interaction of Cognitive Control and Affective Processing**

Research in affective and cognitive neuroscience has begun to identify a common neural substrate involved in moral cognition (Young 2011), in which the prefrontal cortex (PFC) plays a significant role (Zahm, 2006). The PFC has been linked with cognitive control and is crucial for goal-oriented behavior and decision-making. The PFC in primate species is significantly larger in ratio compared to in non-primate species. Furthermore, the white matter (the axonal tracts) in the human anterior PFC is significantly pronounced compared to other primate species, hence suggesting that the cognitive capabilities that are uniquely human may be due to the expanded interconnections of the human brain. The human PFC consists of four subdivisions; lateral prefrontal cortex, frontal pole, medial frontal cortex, and ventromedial prefrontal cortex (vmPFC) including its most ventral part commonly referred to as the orbitofrontal cortex (OFC) (Gazzaniga, Ivry, & Mangun, 2014).
The PFC is consistently elicited by moral sentiment in general, however there is also research suggesting that certain subdivisions are more engaged in some types of moral sentiments than others (Moll et al., 2005). In line with this, several studies confirm the importance of functional connectivity of the vmPFC, OFC, and the amygdala for prosocial behavior and moral cognition (Blair, 2007; Decety & Porges, 2011; Decety, Michalska, & Kinzler, 2012; Heekeren et al., 2005; Moll et al., 2002b). The amygdala enables the individual to learn the goodness and badness of objects and actions and is crucial for stimulus-reinforcement learning (Blair, 2007). Hence, the amygdala is also involved in making affective judgments with emotional characteristics, such as evaluating the aversive or appealing information of a situation, rather than making judgments based on the evaluation of signals elicited by recall of previous emotional life events (Phan et al., 2004). This is further supported by a functional magnetic resonance imaging (fMRI) study in which the participants were to make moral judgments based on written sentences describing agents committing actions containing bodily harm on not. The amygdala showed no greater activation when sentences containing bodily harm were processed as compared to sentences without bodily harm (Heekeren et al., 2005).

The importance of the insula for affective processing has also been highlighted (Decety et al., 2012b). The insula has been linked to the awareness of feelings, as well as the representations of somatic states (Craig, 2009). Hence, both amygdala and insula have been linked to the experience of aversiveness for moral transgressions, which in turn can serve as a guide for the individual to not commit moral transgression.

**Moral dilemmas: Utilitarian and deontological judgements.**

According to Shenhay and Greene (2014), amygdala activation in response to harmful moral transgressions predicts deontological judgments. This is based on an fMRI study in which the integrative roles of the amygdala and the vmPFC when faced with a moral
dilemma were investigated. The moral dilemma depicted to the participants was the trolley problem, hence involving the consideration of performing a harmful action in order to yield the best consequences for as many individuals as possible. Amygdala activation was linked to negative automatic self-reported emotional responses, such as aversiveness, when faced with a moral dilemma. In contrast, the vmPFC was linked with controlled responses favoring utilitarian judgements. Hence, the amygdala provides an affective assessment to the action in question, and the level of activation predict the likelihood of a deontological judgement. In contrast, the vmPFC integrates the amygdala signal with a utilitarian assessment of expected outcomes in order to make a judgement based on the consideration of all competing decisions weights (Shenhav & Greene, 2014). Hence, the amygdala-vmPFC connectivity varies depending on the degree of emotional inputs in the task, being the lowest with pure utilitarian assessments, and the highest with pure emotional assessments (Shenhav & Greene, 2014).

Damage to the vmPFC reliably favors utilitarian judgments due to the amygdala requiring the vmPFC interaction for the making of deontological judgements (Greene, 2015). Thus, the vmPFC is not a center for utilitarian judgements, but rather the necessary integrator which evaluate the affective assessments of the amygdala, (i.e. the affective response to a moral dilemma typically rendering a deontological moral judgement), and hence favors the making of utilitarian judgements. In other words, damage to the vmPFC favors utilitarian judgements, not by damaging a brain region with inherent deontological properties, but by damaging a pathway that is crucial for deontological judgements to prevail (Greene, 2015).

Similarly, in a study by Shenhav and Greene (2010) the same type of moral dilemma as previously mentioned was investigated. However, the quantity of saved lives by performing a prototypical harmful action varied, as well as the probability of them dying without action. The right anterior insula was found to be sensitive to outcome probability (i.e. the probability of the harmful action rendering the desired result of saving lives), which
predicted behavioral sensitivity to this variable. Likewise, the vmPFC and ventral striatum was found to be specifically sensitive to the magnitude of outcome (i.e. group size), in which a similar prediction for behavioral sensitivity for this variable was observed in the ventral striatum. The vmPFC was also found to be involved in the integration of the two variables, hence reflecting a trade-off process between decision weights (magnitude and probability). Thus, Shenhav and Greene (2010) highlight the importance of the frontostriatal pathway in regards of evaluating expected moral value of an action.

**Altruism, cooperation and pro-social behavior.**

The frontostriatal pathway is also involved in altruism and cooperative behavior. Activity in this pathway tracks the value of charitable contributions (Hare, Camerer, Knoepfle, O’Doherty, & Rangel, 2010; Moll et al., 2006), willingness to share resources with others (Zaki & Mitchell, 2011), and cooperation (Rilling et al., 2007), as well as punishing individuals who are insufficiently friendly in their behavior (de Quervain et al., 2004; Singer et al., 2006). Activity in the insula predicts aversion to inequality in the distribution of resources (Hsu et al., 2008). Thus, the brain uses reward signals to motivate cooperative behavior, and negative affective responses to deter uncooperative behavior (Greene, 2015).

Moral judgements that are based on the social behavior of others activates distinct neural networks and regions such as the medial OFC, the temporal pole, and the superior temporal sulcus (STS) (Moll, De Oliveira-Souza, Bramati, & Grafman, 2002a). The medial OFC and the PFC have been shown to be significantly more active when processing prosocial emotions such as guilt, embarrassment, and compassion. And thus, these areas may be essential for the ability to make inferences of possible social outcomes of moral judgments (Moll et al., 2007).

The temporoparietal junction (TPJ) (Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012), as well as medial PFC activity (Rilling et al., 2007; Waytz, Zaki, & Mitchell, 2012)
have both been shown to predict altruistic behavior. Further neural associations with altruism have been found in the default mode network (DMN) (Greene, 2015), which is a network of brain regions that are active when individuals are engaged in internally focused tasks, such as autobiographical memory retrieval, envisioning the future, or adopting the perspectives of others (Buckner, Andrews-Hanna, & Schacter, 2008).

**Empathy and compassion.**

As previously mentioned, three components constitute the three pillars of empathy: the affective component, the motivational component, and the cognitive component (Decety & Cowell, 2015). The affective component refers to the seemingly automatic process that allows for the contagion or sharing of an emotion or an affective state. Emotional sharing or contagion has deep evolutionary roots and is therefore thought of as being hardwired in the brain (Decety & Svetlova, 2012). Several studies have demonstrated that brain regions involved in first-hand experience of pain are also activated when individuals are exposed to facial expressions of pain, or imagine others in pain (Lamm et al., 2011). These regions include the anterior cingulate cortex (ACC), insula, amygdala, somatosensory cortex, as well as frontal- and parietal cortical areas. Notably, the magnitude of the neural activity in these areas correlates stronger with the context in which the nociceptive stimuli appear and with perceived saliency of stimuli, rather than with the intensity or unpleasantness of the nociceptive stimuli (Legrain, Iannetti, Plaghki, & Mouraux, 2011). Furthermore, lack of empathy in children and youths with psychopathic traits have been associated with reduced activity within the neural structures usually implicated in affective responses to the pain or distress of others (ACC, insula, and amygdala) (Cheng, Hung, & Decety, 2012; Marsh et al., 2013). This blunted affective response may contribute to callous disregard for the rights and feelings of others (Decety & Cowell, 2015).
The motivational component refers to the motivation to care for the wellbeing of another individual. For all social mammals, survival and reproduction are dependent on other conspecifics, especially parental care for infant survival and development. The level of care varies for each species, but the underlying neural circuitry for responding to infants seems to be universal (Decety & Cowell, 2015). Converging evidence from animal research (Insel & Young, 2001) and neuroimaging studies on healthy individuals and lesion studies in neurological patients demonstrates that caring for others activates several neural systems and mechanisms such as the PFC including the vmPFC and OFC, as well as the amygdala, and ACC (Shamay-Tsoory, 2009). Furthermore, it also involves the hypothalamic-pituitary-adrenal (HPA) axis, the autonomic nervous system, and brainstem (Decety & Cowell, 2015). The HPA axis is associated with cortisol release in response to a stressful stimulus (Porges 2001; 2007). The autonomic nervous system and brainstem are fundamental to affective experience, emotional expression, facial gestures, vocal communication, and contingent social behavior (Carter, Harris, Porges, 2009). Moreover, endocrine- and hormonal systems are involved in the caring of others (Decety & Cowell, 2015), particularly the hypothalamus in which oxytocin and vasopressin are synthesized (Carter, Harris, & Porges, 2009). Oxytocin facilitates the development of maternal behavior, social bonds and behavioral responses necessary for empathy. It also reduces autonomic and behavioral reactivity to stressful experiences. In contrast, vasopressin is associated with arousal and vigilance (Carter, 2007). These systems regulate bodily states, emotions, as well as social attachment (Shamay-Tsoory, 2009), and seem to borrow aversive signals associated with pain to indicate threats to social relationships (Eisenberger & Lieberman, 2004).

The vmPFC is not necessary for affective responses per se, but functions as an integrator that links concepts with brainstem systems capable of coordinating organism-wide emotional behavior. This integrative process, enabled by the vmPFC, binds together large-
scale networks involved in several functions that are necessary for memory and future projection, self-perception, social cognition, emotion, reward, and autonomic and endocrine function (Roy, Shohamy, & Wager, 2012). Furthermore, the vmPFC has also been linked with empathic concern, in which individual differences predict the magnitude of response in the vmPFC in moral evaluations where guilt is induced (Zahn, de Oliveira-Souza, Bramati, Garrido, & Moll, 2009).

The final component of empathy is the cognitive component, which refers to the ability of affective perspective taking. According to Oswald (1996), affective perspective taking involves the identification and understanding of the feelings of others. Conversely, Oswald (1996) goes on to explain that cognitive perspective taking involves the recognition and understanding of the thoughts of another individual. Affective perspective taking has been linked to social competence and social reasoning (Decety & Cowell, 2015). The neural circuits underpinning affective perspective taking overlap with those supporting ToM and executive functions, thus including brain regions such as the medial frontal cortex, the posterior superior temporal sulcus (pSTS), and the TPJ, which will be discussed in the following section. Furthermore, affective perspective taking activates neural circuits common to those underlying first person experiences, such as the insula, amygdala, and ACC (Jackson et al., 2006; Lamm, Meltzoff, & Decety, 2010; Ruby & Decety, 2004). Additionally, it also activates neural circuits associated with self/other distinction, such as the executive attention network, the frontal cortex including the frontopolar cortex and vmPFC, the right inferior temporal lobe, somatosensory cortex, and the posterior cingulate/precuneus (Decety & Cowell, 2015; Jackson et al., 2006; Lamm et al., 2007; Lamm et al., 2010; Ruby & Decety, 2004).

As previously mentioned, the moral emotion of compassion relies on the integration of social and contextual features. According to Moll et al. (2005), the PFC is associated with
processing situational context (e.g. the orphan child with slim chances of adoption), the STS and anterior temporal cortex have been linked to the context independent social perceptual cues (e.g. distressed facial features of the child) and abstract comprehension of social concepts (e.g. the vulnerability of a child). Furthermore, limbic brain structures, such as the hypothalamus, amygdala, and ventral striatum, are associated with processing central motive states (e.g. experiencing distress, anxiety, social attachment, or hunger) (Moll et al., 2005).

**Perspective taking and theory of mind.**

One of the most primary features of moral evaluation is the one between intentional and accidental harm. The importance of the TPJ, along with other DMN regions have been highlighted for representation and application of this distinction (Greene, 2015). Furthermore, the TPJ has been shown to be specifically sensitive to attempted harms which are morally wrong due to the offender’s mental state, rather than the outcome of the action (Koster-Hale, Saxe, Dungan, & Young, 2013; Young, Cushman, Hauser, & Saxe, 2007).

The TPJ, as well as the pSTS and medial PFC, have also been implicated with ToM and affective perspective taking (Frith & Frith, 2006; Mitchell, 2009). When adopting the perspective of another person, similar neural circuits as the ones underpinning first-person perspectives are activated (Jackson et al., 2006; Lamm et al., 2007; Ruby & Decety, 2004). However, adopting the perspective of another is cognitively demanding and hence requires executive function for additional attentional resources and working memory (Decety & Cowell, 2015). Thus, perspective-taking of another activates previously mentioned neural circuits associated with self/other distinction (Decety & Cowell, 2015; Jackson et al., 2006; Lamm et al., 2007; Lamm et al., 2010; Ruby & Decety, 2004), as well as the right anterior temporal lobe, which is critical for storage and retrieval of social conceptual knowledge and for the experience of moral emotions (De Oliviera-Souza et al., 2015).
Projecting oneself into an aversive situation elicits higher personal distress, which is associated with increased activation of the ACC and amygdala. In contrast, observing the behavioral and emotional reactions of another individual in distress is associated with lower personal distress, higher empathic concern, as well as increased activity in the vmPFC, the executive attention network, and reduced amygdala response (Lamm et al., 2007).

Engagement in moral decision-making when faced with personal and impersonal dilemmas has been associated with increased activation in similar brain regions as those engaged in ToM and perspective taking. Personal dilemmas elicit increased activity in the DMN and the TPJ, as well as the medial PFC, medial parietal cortex, medial frontal gyrus, posterior cingulate gyrus, bilateral STS, and temporal pole (Greene et al., 2001), all of which have also been associated with affection processing (Greene & Haidt, 2002; Kosslyn et al., 1996; Maddock, 1999; Reiman, 1997; Reiman et al., 1997). Greene (2015) goes on to explain that this increased activity in the DMN is to be regarded as due to the relatedness of increased emotional engagement, but not as its proper substrate. The introspective qualities of personal dilemmas elicit especially gripping mental imagery, which may be both a cause and a consequence of their emotional salience.

The dual-process theory.

The interconnection of cognitive control networks and affective processing networks provide neural representative support for the dual-process theory. Impersonal dilemmas, compared to personal dilemmas, elicit greater activity in areas associated with working memory and other controlled cognitive processes, such as the frontoparietal control network, including the vmPFC and dorsolateral prefrontal cortex (dPFC) (Greene et al., 2001; Greene & Haidt, 2002). Moreover, Greene et al. (2001) observed a higher frequency of utilitarian responses to impersonal dilemmas compared to personal dilemmas with the same outcome. This has been linked with greater activation in brain regions associated with socioemotional
processing in personal dilemmas (Greene et al., 2001). Greene (2015) explains further that the utilitarian response to a moral dilemma depends on the explicit cost-benefit reasoning enabled by the frontoparietal control network (Miller & Cohen, 2001), including the dlPFC. Consistent with this, Greene et al. (2004) found that utilitarian responses to impersonal dilemmas, in which this controlled cognitive response tends to dominate, elicit increased dlPFC activity. Likewise, enhanced dlPFC activation was also found in response to personal dilemmas. However, Greene et al., (2004) argues that this increased dlPFC activation is due to the overruling of negative automatic emotional responses in order to make utilitarian judgements, a process which recruits controlled cognitive processes.

Similarly, in an fMRI study by Cushman, Murray, Gordon-mcKeon, Wharton, and Greene (2012) it was found that the tendency to judge harmful actions as morally worse than harmful omissions (i.e. omission effect) is not due to controlled cognition, but rather to the overriding of automatic judgement processes. More specifically, most people would judge it morally worse to actively harm an individual than to passively allow a preventable harm to happen. If this cognitive process were to depend on controlled cognition, then people who exhibit the largest omission effect should then also exhibit the greatest activation in brain regions associated with controlled cognition. However, according to Cushman et al. (2012), increased activation in the frontoparietal control network was associated with condemning harmful omission, hence the opposite relationship was observed. In other words, the enhanced activation in the frontoparietal network was associated with overriding the omission effect. This suggests that the omission effect arises automatically, without the application of controlled cognition (Cushman et al., 2012).

Moreover, several studies investigating the neural correlates of honest behavior provide further support of the dual-process theory. According to Greene (2015), and as previously stated, the dual-process dynamic involves the integration of automatic and
controlled processes in morally laden choices (Greene, 2015). Increased activity in the
frontoparietal control network has been associated with accepting unfair offers, despite their
moral distastefulness (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Tabibnia, Satpute,
& Lieberman, 2008). In a study by Greene and Paxton (2009), participants were given the
opportunity to gain money by being dishonest about their accuracy in predicting the
outcomes of coin-flips. It was found that consistently honest individuals exhibit no increase
in frontoparietal control network activity when refraining from dishonest gains. In contrast,
individuals who behaved dishonestly exhibited increased control-related activity when lying
as well as when forgoing lying. Abe and Greene (2014) argue that these behavioral
differences are due to variations in response characteristics of the frontostriatatal pathway.
Furthermore, Baumgartner, Fischbacher, Feierabend, Lutz, and Fehr (2009) describe a similar
dual-process dynamic in which breaking promises involves increased activity in the
amygdala and the frontoparietal control network.

Additional support for the dual-process theory can be found in studies of patients with
emotion-related deficits. As previously mentioned, damage to the vmPFC favors utilitarian
judgements due to the amygdala requiring the vmPFC interaction for deontological
judgements to prevail (Greene, 2015). Consistent with this, patients with frontotemporal
dementia, who typically demonstrate reduced emotional reactivity, were disproportionately
likely to approve of utilitarian actions in personal moral dilemmas (Mendez, Anderson, &
Shapira, 2005). Similarly, patients with lesions in the vmPFC make up to five times as many
utilitarian judgements in standard high-conflict moral dilemmas (such as performing a
harmful act towards an individual for the greater good) (Ciaramelli, Muccioli, Làdavas, & Di
Pellegrino, 2007; Koenigs et al., 2007). Similar results were found for vmPFC patients in
response to dilemmas pitting the greater good against responsibility towards family, such as
saving the life of a sibling versus the lives of five strangers (Thomas, Croft, & Tranel, 2011).
Furthermore, in response to such utilitarian judgements, patients with vmPFC lesions exhibit correspondingly low levels of physiological stress responses, and thus may be more utilitarian (Moretto, Làdavas, Mattioli, & di Pellegrino, 2010). Decety and Howard (2013) go on to explain that the neural systems that engage in any morally laden scenario are not domain specific. Instead, they support domain-general processing such as intention understanding, affective arousal, attention, and decision making.

**Moral Enhancement**

Moral cognition is malleable and physiological changes in the brain and body can influence moral judgment and behavior (Crockett & Rini, 2015). This section will investigate the possibilities of training-induced plasticity and enhancement of moral cognition.

According to Hursthouse & Pettigrove (2016), a virtue is a morally exemplary and entrenched trait of character. It is a disposition which makes the possessor feel, value, notice, choose, react, and behave in alignment with this virtue. Hence, the possession of a virtue shapes an individual’s personality and character. Similarly, according to Peterson and Seligman (2004), a character strength is a distinguishable way of manifesting a virtue. Hence, the virtue of humanity could be displayed through the character strength of kindness which involves generosity, altruism, care and compassion. Peterson and Seligman (2004) developed the Values in Action (VIA) classification of strengths to describe good character as a fundamental part of optimal human functioning. The VIA classification is comprised of 24 character strengths, subdivided into six categories; cognitive strengths, emotional strengths, interpersonal strengths, civic strengths, strengths of temperance, and strengths of transcendence. According to Peterson and Seligman (2004), these character strengths can be developed by favorable circumstances, as well as by practice.
A study by Kanske et al. (2015) highlights the separation of neural networks associated with affect sharing (empathy) and the understanding of others’ mental states (ToM). According to Kanske et al. (2015), there are distinct neural circuits for the two paradigms, where empathy and affect sharing are, as previously mentioned, associated with neural networks including the insula, whereas ToM has been associated with brain regions such as the TPJ. In an attempt to study these neural networks separately, a novel fMRI paradigm was presented (EmpaToM) which independently manipulated both empathy and ToM. The level of brain activity in one of these networks predicted performance in corresponding paradigm, but not the other (Kanske et al., 2015). This suggests that the distinct brain networks underlying socioaffective and sociocognitive processes can be trained separately by specific trainings targeting either of these paradigms (Kok & Singer, 2017). In the following subsections a few examples of such trainings will be given.

Socioaffective Training

One way to practice and develop character strengths (such as kindness), as well as socioaffective abilities, is through Loving-kindness meditation (LKM). Fredrickson, Cohn, Coffey, Pek, and Finkel (2008) explains that LKM is a technique used to increase warm and caring feelings for others, as well as for the self. In practice, LKM initially involves quiet contemplation over warm and friendly feelings towards another individual who one hold dear. Subsequently, LKM involves extending these cultivated positive thoughts and feelings firstly to the self and then to an expanding circle of others, eventually including all sentient beings (Fredrickson et al., 2008).

In an fMRI study by Lutz, Brefczynski-Lewis, Johnstone, and Davidson (2008), it was found that compassion training, by practicing LKM, was associated with greater activation in the insula and cingulate cortices in response to negative emotional stimuli (sounds of an individual in distress) during meditation compared to when resting.
Furthermore, expert meditators, compared to novices, displayed greater activation in the insula in response to negative sounds than positive or neutral sounds. The enhanced insula-activation was also associated with self-reported intensity of the meditation for expert practitioners as well as novice practitioners. LKM also elicited greater activity in amygdala, the TPJ, and the pSTS in expert meditation practitioners compared to beginners in response to negative, positive and neutral sounds. Hence indicating that long term LKM practitioners develop a greater sensitivity in regards of detection of emotional sounds. Altogether, according to Lutz et al. (2008), these findings suggest that practice of LKM enhances the ability to cultivate positive emotions which can alter the activation of previously linked socioaffective (empathy) and sociocognitive neural networks (ToM) in response to affective stimuli.

In a study by Kang et al. (2014), LKM was also observed to reduce implicit negative attitudes towards stigmatized social groups (Black people and homeless people). The participants were healthy non-Black and non-homeless adults. The LKM practice was divided into two conditions; a 6-week long LKM practice (which contained daily meditation practice 6 days per week), and a 6-week long LKM discussion (in which the participants had one weekly group discussion about received reading materials regard LKM). Reduced implicit interracial and intergroup biases were only observed in the LKM practice group, indicating that the mental expertise to cultivate positive emotion had a greater impact than mere cognitive understanding of them. Furthermore, LKM practice had greater impact on implicit attitudes which are automatically activated, compared to controlled conscious categorization processes. The LKM practice group also displayed a reduction in psychological stress (Kang et al., 2014), which could also render personal benefits such as mood improvement (Fredrickson et al., 2008). Kang et al. (2014) goes on to explain that the observed stress reduction was correlated with reduced bias towards homeless people but not towards Black
people. Thus, relaxations practices solely aimed at stress reduction are unlikely to yield an all-inclusive reduction in implicit biases. Furthermore, increased cognitive control in LKM practitioners was correlated with reduced interracial bias, however there was no significant difference in the two groups, or compared with controls, thus making cognitive control an implausible mediator for this effect (Kang et al., 2014). In contrast, Hunsinger, Livingston, and Isbell (2013), found that LKM practice indeed increased cognitive control, which according to Hunsinger et al. (2013), likely was due to the concentration aspect of LKM. However, the participants’ cognitive control abilities in the study by Hunsinger et al. (2013) were measured immediately after completed LKM practice, hence probably measuring state-like changes in cognitive control. Whereas in the study by Kang et al. (2014), abilities in cognitive control were measured three to seven days after completed LKM course, hence probably measuring trait-like changes in cognitive control. Thus, LKM practice is likely to target socioaffective modes of processing rather than sociocognitive modes of processing (Kang et al., 2014).

However, this affective component of compassion is not entirely straightforward. This since personal distress can lead to avoidance behavior in response to an unpleasant stimulus, such as another’s negative emotional state (Decety & Jackson, 2006). Additionally, empathic concern is an important motivator for altruistic- and helping behavior (Decety, Norman, Berntson, & Cacioppo, 2012a). Thus, Weng et al. (2013) argue that altruistic responses toward another individual’s distress can be strengthened through either reducing personal distress or increasing empathic concern. In a pursuit to investigate this clam, Weng et al. (2013) conducted an fMRI study with healthy adults who underwent either compassion training (COM) or reappraisal training (REP), both intended to train emotion regulation. COM participants cultivated affective abilities by directing compassion towards different targets, hypothesized to increase empathic concern and the urge to alleviate distress. REP
participants practiced active cognitive control by re-interpreting subjectively stressful events, a method hypothesized to reduce negative affect. Altruistic behavior increased after COM training compared to REP training. This effect was associated with altered coordinated activation in the inferior parietal cortex (IPC) and dIPFC, as well as dIPFC connectivity with nucleus accumbens (NAc). Greater activation in the IPC predicted enhanced altruistic behavior, suggesting that the IPC may be a neural indicator for training induced enhanced compassion and altruistic behavior (Weng et al., 2013). Moreover, the IPC has also been associated with mirror neurons (Gallese, Keysers, & Rizzolatti, 2004), and thus may contribute to enhanced simulations of another individual’s distress which may enable possible aversive responses. Hence, the increased coordinated activation of the dIPFC and the IPC may reflect greater sustained attention and goal oriented helping behavior. Furthermore, the NAc has been associated with anticipation of rewards (Knutson, Adams, Fong, & Hommer, 2001), and reappraisal success (reduction in reported emotions) (Wager, Davidson, Hughes, Lindquist, & Ochsner, 2008). Hence, increased dIPFC-NAc connectivity may implicate increased reward value of alleviating distress of another individual in distress, as well as enhanced emotion regulation (Weng et al., 2013).

**Sociocognitive Training**

One way of enhancing sociocognitive abilities is through reading literary fiction, which has been associated with improved empathy and ToM (Oatley, 2016). As previously mentioned, gripping mental imagery is correlated with emotional engagement (Greene, 2015). Similarly, increased empathy has been correlated with vividness of imagery (Johnson, Cushman, Borden, & McCune, 2013). Furthermore, brain regions associated with narrative comprehension overlap to a large extent with brain regions such as the mPFC, pSTS, and TPJ (Mar, 2011), which also have been linked to ToM (Decety & Cowell, 2015; Frith & Frith, 2006; Mitchell, 2009).
In a study by Speer, Reynolds, Swallow, and Zacks (2009) it was found that brain regions involved in spatial navigation, manual manipulation of objects, and goal-directed activity, such as the premotor cortex, increased in activity when the reader processed a fictional situation involving those specific aspect (Speer et al., 2009). Thus, fiction seems to be a simulation of the self in interaction (Oatley, 2016).

The improvement of empathy and ToM induced by reading literary fiction may be due to the promotion of subject matter in literary fiction (i.e. understanding people and their intentions), compared to non-fiction which promotes expertise in its specific content (e.g. astronomy), as well as popular fiction (Oatley, 2016). The distinction between literary fiction and popular fiction lies in the complexity of the character, which allows for the reader to have more refined and cultivated inferences regarding their motives and emotions (Koopman & Hakemulder, 2015), which in turn can aid in the breaking of stereotypes (Oatley, 2016). Hence, exposure to fiction may reduce racial biases. According to Johnson, Huffman, and Jasper (2014), this effect was observed when exposure to a fictional story about a counter-stereotypical Muslim woman reduced racial biases between Arab and Caucasian faces. This may be due to fictional characters enabling the imagination of what it might be like to be in other peoples’ situations (Johnson et al., 2014). Moreover, literary fiction enables individuals to emotionally engage in an increased number of life-circumstances than what would be likely to encounter in real life. Hence, giving us insight in, and experience of, other peoples’ lives and perspectives, which can aid in our understanding and concern for others’ well-being (Oatley, 2016).

**Sociocognitive and Socioaffective Training with Interactive Aspects**

The previously mentioned trainings are all intended to enhance prosocial behavior by cultivating skills such as empathy, compassion, and greater understanding of others. However, the performance of these activities is usually carried out as a solitary practice.
In the pursuit of investigating the explicit psychological effects of implementing an interactive practice (in which two individuals verbally share their thoughts, feelings, and experiences with one another, i.e. contemplative dyads), Valk et al. (2017) conducted a large-scale 9-month longitudinal study of secularized contemplative practices. Furthermore, Valk et al. (2017) implemented previously mentioned research supporting the view of distinct neural networks underlying sociocognitive- and socioaffective abilities in order to train each paradigm separately, as well as mindfulness-based present-moment-focused attention. These trainings were divided into three modules; the presence module, the affect module, and the perspective module, which contained interventions aimed at improving the specific affective and cognitive functions of compassion and ToM (Valk et al., 2017).

The presence module was designed to train attention and interoceptive awareness through breathing meditation and body scan interventions, hence lacking the dyadic element. The breathing meditation involved focused attention on the breathing and the practice of refocusing wandering attention back to the breathing. The basic instructions for the body scan was to systemically direct focus on various body parts and observing the occurring sensations. Additional interventions also involved walking meditation with focused attention to auditory- and visual stimuli (Valk et al., 2017).

The affective module was designed to train care, compassion, gratitude, prosocial motivation, and the capability to deal with difficult emotions through the interventions of affect dyads and LKM. In the affect dyads, two individuals take turns in listening and expressing feelings and bodily sensations during positive and negative experiences during the day, whilst the other individual practices silent but active empathic listening. Additional elements in the affect module focused on cultivating an accepting, kind and compassionate attitude for the self and others, including forgiveness meditation and development of self-compassion (Valk et al., 2017).
The perspective module was designed to train metacognition, perspective-taking on self, and perspective-taking on others through the interventions of perspective dyads and observing-thoughts meditation (OTM). In the perspective dyads, two individuals alternate to describe a recent event from the perspective of a randomly assigned perspective or “inner part” (e.g. “the loving mother”) (Kok & Singer, 2017, p.127) whilst the counterpart practices mindful listening to infer what perspective is currently being presented. In the OTM intervention, the participants practice labelling mental events and to observe them without automatically reacting. An additional part of the perspective module included different interventions in which the participants took on the perspective of someone who they had daily struggles with. Furthermore, the participants were randomly assigned into three training cohorts (TC), TC1 and TC2 each started with the prescience module, followed by the affect module and perspective module in alternating order, whereas TC3 consisted only of the 3-month long affect module (Valk et al., 2017).

According to Valk et al. (2017), these interventions resulted in an observed increase in cortical thickness in the right PFC extending to ACC, as well as in the bilateral occipital regions extending to inferior temporal cortices for both TC1 and TC2 in the presence module. Furthermore, for the affect module, which was intended to train socioaffective skills, increased cortical thickness was observed primarily in a network extending from the right supramarginal gyrus (SMG) to the insular-opercular areas and the dLPFC, left mid/posterior cingulate cortex, as well as bilateral parahippocampal areas. Lastly, the perspective module, which cultivated sociocognitive abilities, was correlated with an increase in cortical thickness in the left occipital regions, the left ventrolateral PFC, right middle temporal gyrus. See figure 1 for an illustration of the affected brain areas for each module.
MORAL COGNITION AND ITS NEURAL CORRELATES: POSSIBILITIES FOR ENHANCEMENT OF MORAL COGNITION AND BEHAVIOR

Figure 1. Module-specific training-related cortical thickness increases for the presence, affective, and perspective module. For the presence module, affected areas (marked in yellow) include; right PFC, ACC, bilateral occipital regions, and inferior temporal cortices. For the affect module, affected areas (marked in red) include; right SMG, insular-opercular areas, dlPFC, left mid/posterior cingulate cortex, and bilateral parahippocampal areas. For the perspective module, affected areas (marked in green) include; left occipital regions, left ventrolateral PFC, and right middle temporal gyrus. Reprinted from “Structural plasticity of the social brain: Differential change after socio-affective and cognitive mental training” by S. L. Valk, B. C. Bernhardt, F. M. Trautwein, A. Böckler, P. Kanske, N. Guizard, … T. Singer, 2017, Science Advances, 3(10), 1-11. Reprinted with permission.

In a related but yet unpublished study, Trautwein, Kanske, Böckler-Raettig, and Singer (2017) observed performance enhancement in a task intended to assess executive control of attention after completed presence training. Similarly, socioaffective training induced enhanced compassion ratings regarding emotionally neutral or distressing video narratives. Lastly, sociocognitive training was correlated with enhanced performance in an interactive video task intended to assess ToM (Trautwein et al., 2017). Individual differences in these performance enhancements correlated with training-induced cortical thickening in the same individuals. Performance enhancement in attention after presence training correlated with increased cortical thickness in the left middle temporal regions. Furthermore, enhanced compassion rating after completed socioaffective training correlated with increased thickness.
MORAL COGNITION AND ITS NEURAL CORRELATES: POSSIBILITIES FOR ENHANCEMENT OF MORAL COGNITION AND BEHAVIOR

in the right insular regions extending to the temporal pole. Lastly, increase in ToM performance correlated with cortical thickness increase of the left parietal regions and the right TPJ after sociocognitive training (Valk et al., 2017). See figure 2 for an illustration of the affected brain areas for each module.

**Figure 2.** Behavioral modulation of training related cortical thickness change. For the presence module, affected areas (marked in yellow) include the left middle temporal regions. For the affect module, affected areas (marked in red) include the right insular regions extending to the temporal pole. For the perspective module, affected areas (marked in green) include the left parietal regions and right TPJ Reprinted from "Structural plasticity of the social brain: Differential change after socio-affective and cognitive mental training" by S. L. Valk, B. C. Bernhardt, F. M. Trautwein, A. Böckler, P. Kanske, N. Guizard, … T. Singer, 2017, Science Advances, 3(10), 1-11. Reprinted with permission.

Furthermore, Valk et al. (2017) also observed that the training-induced structural neurological change correlated with socioaffective and sociocognitive functional activation during previously mentioned tasks. However, this was not observed for the attention task.

Altogether, the brain regions and networks which correlated with training-induced cortical thickening and enhancement of socioaffective and sociocognitive abilities largely correlated with the previously linked brain regions and networks with empathy, compassion, ToM, and executive attention control. Hence, according to Valk et al. (2017), providing support for structural plasticity of moral cognition in healthy adults. Thus, promoting mental
training by utilizing evidence-based interventions aimed at cultivating prosocial behavior and cooperation (Valk et al., 2017).

Discussion

The aim of this essay was to provide an overview of some key theories and frameworks regarding moral cognition and its neural correlates, in order to examine the possibilities of enhancement of moral cognition. This section will compile the presented information in previous sections in order to highlight and discuss correlations, as well as suggest future research.

There are several propositional models and theories of morality, however the distinguishing factor between them is the varying emphasis put on their respective affective and cognitive components. The rationalist model emphasizes reason and reflection for the making of moral judgements, whereas affective states are mere inputs to the reasoning process. In contrast, the social intuitionist model emphasizes the importance of moral intuitions which require no ratiocination. Here, any justification or rational processing regarding moral judgements, is an ex post facto process (Haidt, 2001). However, according to Greene et al. (2008) and the dual-process theory, it is the integration of the initial automatic emotional response and the slower cognitive response to a moral dilemma which constitutes moral judgements.

According to Greene (2015) there are no specific moral faculty in the brain from which moral cognition arises. Instead, it relies on the functional integration of multiple neural mechanisms and cognitive systems (Greene, 2015). As previously mentioned, empathy has primarily been correlated with increased activity in the ACC, insula, amygdala, somatosensory cortex, as well as frontal-and parietal cortical areas in response to observing another individual in distress or pain (Lamm et al., 2007; Lamm et al., 2011; Legrain, Iannetti, Plaghki, & Mouraux, 2011). Similarly, compassion has also been primarily
associated with the frontal- and parietal cortical areas, amygdala, ACC, as well as the STS, HPA axis, and autonomic nervous system (Decety & Cowell, 2015; Moll et al., 2005; Shamay-Tsoory 2009).

Interestingly, similar brain regions and neural networks that have been associated with empathy and compassion were consistently involved in socioaffective training and enhancement. According to Lutz et al. (2008), compassion training (by practicing LKM) increased neural activity in the insula, cingulate cortices, and amygdala in response to sounds of others’ distress. Furthermore, increased activity in these brain regions was also associated with performance enhancements in picking up interpersonal affective cues (Lutz et al., 2008). Correspondingly, LKM and affect dyads induced an increase in cortical thickness in the insular regions, cingulate cortices, as well as executive neural regions (such as the dlPFC), and parietal areas (such as the SMG). Here, performance enhancement was correlated with cortical thickness increase in the right insular regions extending to the temporal pole (Valk et al., 2017). Moreover, frontal and parietal regions have also been associated with enhanced altruistic behavior induced by compassion training. More specifically, compassion training has been observed to alter the coordinated activity in the dlPFC and the IPC, where increased activity in the IPC predict altruistic behavior (Weng et al., 2013). The association of the IPC and mirror neurons has also been highlighted (Gallese et al., 2004), which could be argued to possibly strengthen the motivational component of empathy and thus enable altruistic behavior.

In summary, these combined results support the involvement of the insula, cingulate cortices, and cognitive control networks in reflecting and integrating affective states and emotional cues from internal and external sources. Hence, the insula, cingulate cortices, and cognitive control networks are crucial for the processing of social emotions such as empathy and compassion. Furthermore, these results also support the possibility of adult neural
plasticity and enhancement of socioaffective neural networks, as well as performance enhancement in moral social interactions induced by socioaffective training (Valk et al. 20017).

The sociocognitive component of moral cognition (ToM) has been primarily associated with the mPFC, pSTS, medial parietal cortex, and the TPJ (Decety & Cowell, 2015; Frith & Frith, 2006; Mitchell, 2009), which also overlap with brain regions associated with narrative comprehension (Mar, 2011). Moreover, ToM also primarily activates neural networks linked with self/other distinction. These areas include frontal cortical regions (such as the frontopolar cortex and vmPFC), parietal regions (such as the somatosensory cortex, the posterior cingulate cortex, and the right inferior temporal lobe) (Decety & Cowell, 2015; Jackson et al., 2006; Lamm et al., 2007; Lamm et al., 2010; Ruby & Decety, 2004).

Alike the overlap of brain regions elicited by socioaffective training and the neural correlates of empathy and compassion, sociocognitive training was observed to activate and affect similar brain areas and neural networks as the previously stated neural correlates of ToM. As such, OTM and perspective dyads were observed to induce cortical thickening of the left parietal regions, left ventrolateral PFC, the right middle temporal gyrus, and right TPJ. Here, increased ToM performance was primarily correlated with cortical thickness increase of the left parietal regions and the right TPJ (Valk et al., 2017). Additionally, enhanced activity in the TPJ and pSTS has been associated with LKM training and expertise (Lutz et al., 2008).

Similar brain regions which have been previously associated with execution of controlled cognitive processes and attention, such as the frontoparietal control network, the vmPFC, and dlPFC cortex (Greene et al., 2001; Greene & Haidt, 2002) were also consistently involved in sociocognitive training and enhancement. According to Valk et al. (2017), presence training interventions induced increased cortical thickness in networks including the
right PFC extending to ACC, as well as in the bilateral occipital regions extending to inferior and middle temporal cortices. Here, performance enhancement in attention correlated with increased cortical thickness in the middle temporal regions.

Furthermore, the structural neurological change induced by socioaffective and sociocognitive training correlated with socioaffective and sociocognitive functional activation (Valk et al., 2017). Functional activation was measured during tasks intended to assess performance in attention, compassion, and theory of mind (Trautwein et al., 2017). However, overlapping functional activation and increased cortical thickness was not observed for the attention task, which Valk et al. (2017) argues may not have successfully targeted the entirety of the presence module. This inadequacy in assessing the entirety of the modules could also be argued to be present for the performance enhancement measurements. Valk et al. (2017) goes on to explain that in each training module, a broad array of skills was trained, going above and beyond the behavioral change in for example compassion for affect (which was measured in the previously mentioned compassion task regarding emotionally neutral or distressing video narratives). Thus, overall change, which represent many changes in a broader spectrum of affect-related behaviors, does not necessarily directly relate to changes in compassion (Valk et al., 2017). Hence, it could also be argued that this incommodes the creation of a task intended to measure performance enhancement in all the skills trained in each module.

In conclusion, this essay has addressed the philosophical and psychological aspects of the key concepts in moral cognition, such as empathy, compassion, and ToM. Furthermore, it has addressed the current research on the neural correlates underpinning these concepts, and correlated this with recent research on moral enhancement. Hence, this essay has bridged the gaps between the theoretical aspects of moral cognition, the neural correlates that enable
these concepts and abilities, and the possibility to enhance these abilities by specific trainings and interventions.

Current research provide support for the distinction between the neural networks that underlie socioaffective and sociocognitive processes. Hence, similarly to physical training which may target specific muscle groups, there are also possibilities to target and train socioaffective and sociocognitive abilities separately. Furthermore, just as muscles react to targeted training with increased strength and physical volume, presented research have suggested that training specific neural networks may yield performance enhancement in the targeted ability, as well as increased cortical volume in corresponding brain regions and networks. Unlike physical training, in which the results only render benefits to the individual, enhancement in moral abilities, such as sociocognitive and socioaffective abilities, may also extend acquired benefits beyond the individual. This since, as previously stated, increased sensitivity to emotional stimuli through compassion training can increase the ability to pick up emotional cues, as well as strengthen the urge and motivation to alleviate others’ distress. Furthermore, practicing and strengthening ToM allows for a better understanding of the perspective of others, which can reduce biases between individuals or societal groups.

For the future, further research on the effects of sociocognitive and socioaffective training on moral judgements is needed. More specifically, investigating further the neural correlates of deontological judgements and utilitarian judgments, and whether compassion training or ToM practice may affect these judgements, perhaps favoring deontological judgements. Thus, applied to the footbridge dilemma and trolley problem, would enhanced moral cognition (i.e. enhanced empathy, compassion, and ToM) make an individual feel more inclined to save a greater number of people, or perhaps experience more aversion towards pushing the stranger or pulling the switch?
References


MORAL COGNITION AND ITS NEURAL CORRELATES: POSSIBILITIES FOR ENHANCEMENT OF MORAL COGNITION AND BEHAVIOR


